PLOS ONE

RESEARCH ARTICLE

Sizing the Problem of Improving Discovery and Access to NIH-Funded Data: A Preliminary Study

Kevin B. Read[1☯]*, Jerry R. Sheehan[2☯], Michael F. Huerta[2☯], Lou S. Knecht[2☯], James G. Mork[2☯], Betsy L. Humphreys[2☯], NIH Big Data Annotator Group[3¶]

1 Medical Library, NYU Langone Medical Center, New York, New York, United States of America, **2** National Library of Medicine, National Institutes of Health, Bethesda, Maryland, United States of America, **3** National Institutes of Health, Bethesda, Maryland, United States of America

☯ These authors contributed equally to this work.
¶ Membership of the NIH Big Data Annotator Group is listed in the Acknowledgments.
* kevin.read@nyumc.org

OPEN ACCESS

Citation: Read KB, Sheehan JR, Huerta MF, Knecht LS, Mork JG, Humphreys BL, et al. (2015) Sizing the Problem of Improving Discovery and Access to NIH-Funded Data: A Preliminary Study. PLoS ONE 10(7): e0132735. doi:10.1371/journal.pone.0132735

Editor: Vincent Larivière, Université de Montréal, CANADA

Received: January 8, 2015

Accepted: June 17, 2015

Published: July 24, 2015

Data Availability Statement: The data analysis file and all annotator data files are available in the Figshare repository /m9.figshare.1285515. Read K. (2015). Sizing the Problem of Improving Discovery and Access to NIH-funded Data: A Preliminary Study (Datasets). Figshare. Available: http://dx.doi.org/10.6084/m9.figshare.1285515.

Funding: This research was supported by the Intramural Research Program of the U.S. National Institutes of Health, National Library of Medicine (NLM) and in part by an appointment to the NLM Associate Fellowship Program sponsored by the

Abstract

Objective

This study informs efforts to improve the discoverability of and access to biomedical datasets by providing a preliminary estimate of the number and type of datasets generated annually by research funded by the U.S. National Institutes of Health (NIH). It focuses on those datasets that are "invisible" or not deposited in a known repository.

Methods

We analyzed NIH-funded journal articles that were published in 2011, cited in PubMed and deposited in PubMed Central (PMC) to identify those that indicate data were submitted to a known repository. After excluding those articles, we analyzed a random sample of the remaining articles to estimate how many and what types of invisible datasets were used in each article.

Results

About 12% of the articles explicitly mention deposition of datasets in recognized repositories, leaving 88% that are invisible datasets. Among articles with invisible datasets, we found an average of 2.9 to 3.4 datasets, suggesting there were approximately 200,000 to 235,000 invisible datasets generated from NIH-funded research published in 2011. Approximately 87% of the invisible datasets consist of data newly collected for the research reported; 13% reflect reuse of existing data. More than 50% of the datasets were derived from live human or non-human animal subjects.

Conclusion

In addition to providing a rough estimate of the total number of datasets produced per year by NIH-funded researchers, this study identifies additional issues that must be addressed to

National Library of Medicine and administered by the Oak Ridge Institute for Science and Education.

Competing Interests: The authors have declared that no competing interests exist.

improve the discoverability of and access to biomedical research data: the definition of a "dataset," determination of which (if any) data are valuable for archiving and preservation, and better methods for estimating the number of datasets of interest. Lack of consensus amongst annotators about the number of datasets in a given article reinforces the need for a principled way of thinking about how to identify and characterize biomedical datasets.

Introduction

Biomedical research is becoming increasingly data-centric. The proliferation of low-cost methods for whole genome sequencing, growing use of functional magnetic resonance imaging (fMRI) and other imaging modalities, and more widespread availability of clinical data in electronic health records (EHRs) are among the factors enabling biomedical researchers to generate and make use of increasing volumes of digital data in their research. Growth in the availability of biomedical data is, in turn, generating growing interest in improving the management and utilization of the many types of data (e.g., genomic, imaging, behavioral, clinical, exposure) that are used in biomedical research.

Improved management of biomedical research data—or any scientific data—can have many benefits. Fundamentally, improved management of scientific data is essential to the preservation of the scientific record, of which data are a growing part. It is also the basis for improved sharing of data, for example, enabling other researchers to have access to previously collected data. Data sharing can improve the quality and efficiency of research by allowing researchers to verify and validate prior research findings, to conduct research that combines previously collected data with newly collected data, and to compare the results of related research studies more easily [1, 2].

Around the world, governments, research funding organizations, and investigators are actively pursuing better management of, and access to scientific research data [3]. The European Commission, European Research Council and Canadian Institute of Health Research have all established policies for research data [4–6]. In the United States, a February 2013 memorandum from the White House Office of Science and Technology Policy directed all U.S. federal science agencies that spend more than $100 million per year on research and development to develop plans to increase public access to digital data resulting from research funded by those agencies [7]. The U.S. Department of Health and Human Services issued its plans in February 2015 [8].

The U.S. National Institutes of Health (NIH), part of the Department of Health and Human Services, is the world's largest funder of biomedical research. It invests approximately $30 billion per year in biomedical research, most of which is expended through competitive grants to more than 300,000 researchers at universities, medical schools, and other research institutions in every U.S. state and around the world [9]. The NIH Big Data to Knowledge (BD2K) initiative launched in 2013 aims to "enable biomedical scientists to capitalize more fully on the Big Data being generated by those research communities" [10]. One goal of BD2K is to develop effective and efficient mechanisms to enable the identification of, access to, and citation for biomedical data, bringing more data into the ecosystem of science and scholarship [11].

An important step in designing, developing, and implementing mechanisms to discover, access, and cite the biomedical data used in NIH-funded research is to characterize the number of new datasets generated annually by NIH-funded researchers, the types of data created, and the frequency of reuse of existing data. Of particular interest are "invisible" datasets—datasets

that are not currently stored and made accessible via well-known, publicly accessible data repositories. Previous studies have estimated how much data is shared by analyzing a set number of journals [12, 13], performed analyses on how often specific datasets were cited in the literature [14, 15], and used complex algorithms to estimate the entire universe of data for a specific discipline [16]. While it has been shown that it is possible to make some estimates of the types of data that are currently deposited in known repositories, it is more challenging to estimate the number of datasets that are *not* publicly or systematically registered, deposited, or archived. Arguably, such datasets should be a primary focus of any effort to improve the discoverability and reuse of data because they are less discoverable and accessible than data deposited in a known repository.

We conducted a study to develop a preliminary estimate of the annual volume and types of datasets generated by NIH-funded researchers. This study was undertaken to inform initial NIH efforts to improve the discoverability of and access to biomedical datasets. For the purpose of this study, a dataset was defined as any collection of data (e.g., different type of measurement) that was generated or reused to inform the results described in an article.

Methods

Our approach to characterizing biomedical research datasets relied on an examination of datasets that are used or generated in the course of research that is reported in published journal literature. This approach misses datasets that are collected as part of a research project but are not reported in a publication. While little is known about the full extent of non-publication in biomedical research, recent work indicates that as long as four years after study completion, the results from approximately one-third of clinical trials registered in ClinicalTrials.gov remains unpublished [17]. There is also evidence that the availability and discoverability of research datasets declines rapidly with age [18]. To the extent that discovery of datasets may be enabled by linking data to associated journal articles, our approach was a reasonable first step toward quantification and characterization. We further restrict our analysis to datasets generated by NIH-funded research. While this does not represent all of biomedical research, it is research that is subject to U.S. policies that require expanded data sharing. This sample also covers a broad spectrum of biomedical research types, from basic to clinical research across a wide range of diseases, conditions, and systems and therefore was a good starting point for analysis. For clarity, the process taken to identify NIH-funded datasets via the published journal literature described below is also illustrated in Fig 1.

Identifying articles with datasets deposited in known repositories

To estimate the number of datasets generated annually from NIH-funded research and identify datasets that are not stored in a known repository, our analysis focused on NIH-funded articles that were published in 2011. These articles represented the most current complete set of articles for a given year at the time of our study. To retrieve these articles, we searched PubMed using the strategy illustrated below (Table 1-a), which retrieved citations from articles published in 2011 that acknowledged research funding support from NIH (step 1a, Fig 1). Use of PubMed's Publication Type [PT] and Grant [GR] search tags enabled us to focus the search on citations that received NIH funding. This search identified almost 120,000 citations. We further limited the search to include those citations indexed in MEDLINE (using the MEDLINE [sb] subset search tag). This step focuses the analysis on citations that have been fully indexed and contain additional information to indicate whether datasets used in the summarized research were deposited in a known data repository. This search (Table 1-b) retrieved more than 113,000

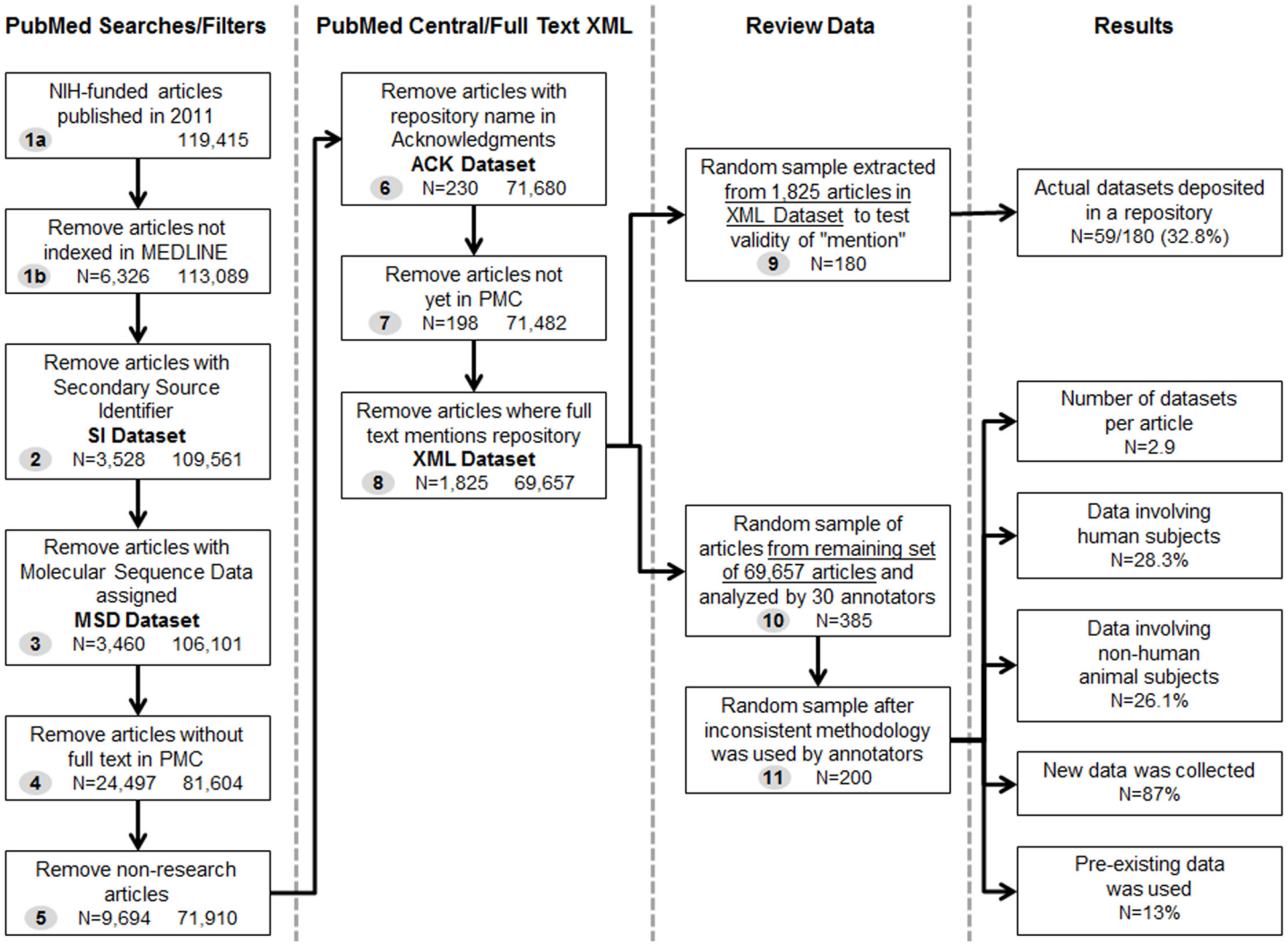


Fig 1. Diagram of process taken to identify NIH-funded datasets via the published journal literature (Including Results).

doi:10.1371/journal.pone.0132735.g001

articles. All bolded text in Tables 1–5 indicate search terms that were progressively added to the search string (step 1b, Fig 1).

From this set of articles, we identified those that indicated when authors shared their data in a specific repository. This process began by searching for articles that had a Secondary Source Identifier [SI][19]; this identifier indicates when the author of an article has deposited his/her data in one of the specific repositories that are recognized in MEDLINE/PubMed. The repositories that can be designated in the SI field include, but are not limited to: ClinicalTrials.gov, PubChem, Johns Hopkins University Genome Data Bank, Gene Expression Omnibus,

Table 1. PubMed searches identifying articles with funding support from the NIH.

a)	2011 [dp] AND (NIH [gr] OR Research Support, N.I.H., Extramural [pt] OR Research Support, N.I.H., Intramural [pt])	119,415
b)	2011 [dp] AND (NIH [gr] OR Research Support, N.I.H., Extramural [pt] OR Research Support, N.I.H., Intramural [pt]) **AND medline [sb]**	113,089

doi:10.1371/journal.pone.0132735.t001

Table 2. PubMed searches identifying when datasets were deposited in certain repositories (SI dataset).

2011 [dp] AND (NIH [gr] OR Research Support, N.I.H., Extramural [pt] OR Research Support, N.I.H., Intramural [pt]) AND medline [sb] **AND (GDB [si] OR GENBANK [si] OR OMIM [si] OR PDB [si] OR PIR [si] OR RefSeq [si] OR SWISSPROT [si] OR ClinicalTrials.gov [si] OR ISRCTN [si] OR GEO [si] OR PubChem-Substance [si] OR PubChem-Compound [si] OR PubChem-BioAssay [si])**	3528

doi:10.1371/journal.pone.0132735.t002

Table 3. PubMed search identifying articles with the "Molecular Sequence Data" MeSH Heading (MSD dataset).

2011 [dp] AND (NIH [gr] OR Research Support, N.I.H., Extramural [pt] OR Research Support, N.I.H., Intramural [pt]) AND medline [sb] **NOT** (GDB [si] OR GENBANK [si] OR OMIM [si] OR PDB [si] OR PIR [si] OR RefSeq [si] OR SWISSPROT [si] OR ClinicalTrials.gov [si] OR ISRCTN [si] OR GEO [si] OR PubChem-Substance [si] OR PubChem-Compound [si] OR PubChem-BioAssay [si]) **AND molecular sequence data [mh:noexp]**	3460

doi:10.1371/journal.pone.0132735.t003

Table 4. PubMed search identifying articles in PMC.

2011 [dp] AND (NIH [gr] OR Research Support, N.I.H., Extramural [pt] OR Research Support, N.I.H., Intramural [pt]) AND medline [sb] **NOT** (GDB [si] OR GENBANK [si] OR OMIM [si] OR PDB [si] OR PIR [si] OR RefSeq [si] OR SWISSPROT [si] OR ClinicalTrials.gov [si] OR ISRCTN [si] OR GEO [si] OR PubChem-Substance [si] OR PubChem-Compound [si] OR PubChem-BioAssay [si]) **NOT** molecular sequence data [mh:noexp] **AND pubmed pmc all[sb]**	81604

doi:10.1371/journal.pone.0132735.t004

Table 5. Removal of articles that were not considered "research".

2011 [dp] AND (NIH [gr] OR Research Support, N.I.H., Extramural [pt] OR Research Support, N.I.H., Intramural [pt]) AND medline [sb] NOT (GDB [si] OR GENBANK [si] OR OMIM [si] OR PDB [si] OR PIR [si] OR RefSeq [si] OR SWISSPROT [si] OR ClinicalTrials.gov [si] OR ISRCTN [si] OR GEO [si] OR PubChem-Substance [si] OR PubChem-Compound [si] OR PubChem-BioAssay [si]) NOT molecular sequence data [mh:noexp] AND pubmed pmc all[sb] **NOT review [pt] NOT letter [pt] NOT news [pt] NOT editorial [pt]**	71910

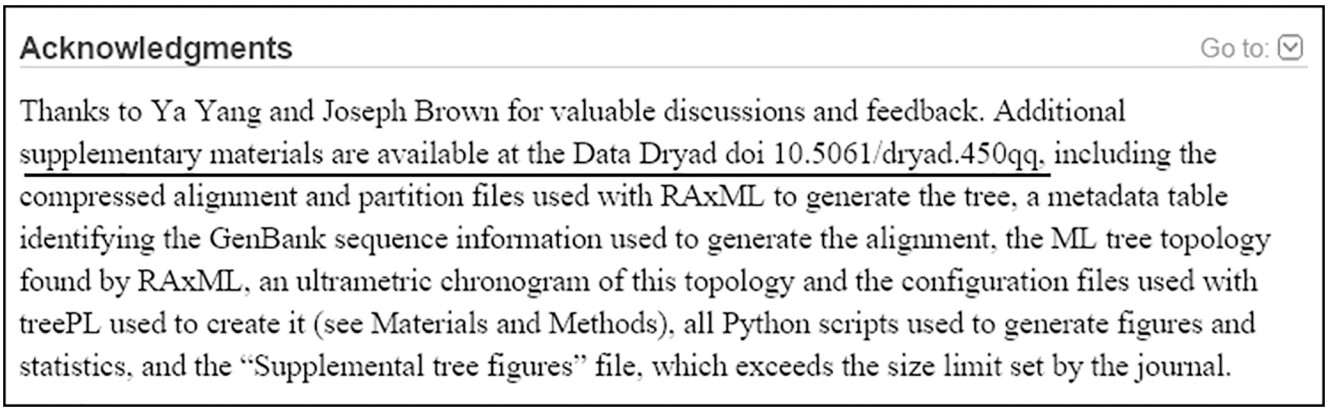
doi:10.1371/journal.pone.0132735.t005

GenBank, ISRCTN Register, Mendelian Inheritance in Man, Protein Data Bank, Protein Identification Resource, Reference Sequence, and SWISSPROT Protein Sequence Database. The aforementioned repositories provide evidence of how many articles acknowledge data deposition in each of these locations within a given year (Table 2). This step identified 3,528 (**SI dataset**) articles that had deposited data into one of the listed repositories (step 2, Fig 1).

We removed the **SI dataset** articles from our article set and searched the remaining articles for the Medical Subject Heading (MeSH) "Molecular Sequence Data." Citations tagged with this MeSH heading are those for which data are likely to be deposited in GenBank or an equivalent repository. For the year 2011 this search identified 3,460 (**MSD dataset**) articles that provided indication that data should have been deposited in a repository (Table 3; step 3, Fig 1).

We removed the **MSD dataset** articles from the sample and searched the remaining articles for those with full-text available in PubMed Central (PMC) using the [sb] search tag. This allowed us to conduct further analysis on information that would be provided only in the full-text of an article (as opposed to the MEDLINE citation) (Table 4; step 4, Fig 1).

We reduced the set of articles by identifying and removing all non-research articles, meaning those with the MEDLINE publication type [PT] of review, editorial, news, and letter. This



Acknowledgments

Go to:

Thanks to Ya Yang and Joseph Brown for valuable discussions and feedback. Additional supplementary materials are available at the Data Dryad doi 10.5061/dryad.450qq, including the compressed alignment and partition files used with RAxML to generate the tree, a metadata table identifying the GenBank sequence information used to generate the alignment, the ML tree topology found by RAxML, an ultrametric chronogram of this topology and the configuration files used with treePL used to create it (see Materials and Methods), all Python scripts used to generate figures and statistics, and the "Supplemental tree figures" file, which exceeds the size limit set by the journal.

Fig 2. Example of the PubMed Central Acknowledgments where the authors have indicated the deposit of data in a specific repository; PMCID: PMC4085032.

doi:10.1371/journal.pone.0132735.g002

step created a sample that included only full-text research articles with MEDLINE records that did not mention depositing data into a repository (Table 5). This process resulted in a sample of 71,910 articles (step 5, Fig 1).

We then examined the articles to identify those that mention the sharing of their data in the acknowledgments section of an article, using the Acknowledgements search field [20] of PMC (Fig 2) [21]. The Acknowledgments section of a full-text article is often used to indicate when data have been shared in a specific repository. We selected the NIH Data Sharing Repositories Web page [22] as our gold standard to gather a list of NIH-specific data repositories, and used keyword variations and acronyms (e.g., Gene Expression Omnibus, GEO, Protein Data Bank, PDB) to search each repository in the Acknowledgments field in PMC with the [ack] search tag for the year 2011. Additionally, the terms "DataCite" and "Dryad" were added to the strategy, seeking occurrences in any PMC search field, because they are well-known resources for discovery of scientific data, including data referenced in scientific journal articles.

This search identified 814 (**ACK dataset**) articles that mentioned one or more of the recognized repositories. After accounting for overlap with the **SI dataset** and **MSD dataset**, we removed 230 articles in the **ACK dataset** from our sample set, leaving us with 71,680 articles that made no mention that their data were deposited in a known repository (step 6, Fig 1). Of these articles, 198 were not yet available in PMC at the time of our study, so they were removed from the sample, leaving 71,482 articles (step 7, Fig 1).

The final procedure used to identify articles that mention the deposit of data was to scan for the same keyword variations and acronyms from the 45 NIH data repositories within the XML full-text data for the remaining articles [23]. This step aimed to fill in any gaps from the two previous strategies and to search beyond the scope of the Acknowledgments field in PMC to find additional mentions of data repositories. It was only possible to perform this search on 10,418 articles for which full-text XML was available via the PMC Open Access Subset [24]; the PMC Open Access Subset includes articles that are still protected by copyright, but are made available via a Creative Commons or similar license that provides for more liberal distribution and reuse of the copyrighted work. This method identified 1,825 articles (**XML dataset**) in total that mentioned a data repository (step 8, Fig 1). We removed these articles from the sample leaving a total number of 69,657 NIH-funded articles that contained "invisible" datasets (Table 6).

The mention of a dataset or repository in the body of the full-text does not necessarily mean that the data are deposited in the repository; it confirms only the presence of the term(s) in the

Table 6. Breakdown for subtraction of articles that mention the deposit of data.

Procedure taken	Articles identified	Articles remaining
1a. NIH-funded articles for 2011 in PubMed	–	119,415
1b. NIH-funded articles for 2011 indexed for MEDLINE	6,326	113,089
2. Articles with repository in [SI] field **(SI dataset)**	3,528	109,561
3. Articles with Molecular Sequence Data MeSH Heading **(MSD dataset)**	3,460	106,101
4. PubMed cited articles not available in PMC	24,497	81,604
5. Non-research articles	9,694	71,910
6. Articles with repository in PMC Acknowledgements **(ACK dataset)**	230	71,680
7. Additional articles not available in PMC	198	71,482
8. Articles with repository in full-text XML (of 10,418 searched) **(XML dataset)**	1,825	69,657
Total remaining articles used for subsequent analysis	–	**69, 657**

doi:10.1371/journal.pone.0132735.t006

article, not the context. To estimate the frequency with which a mention of a repository corresponds to the actual deposit of a dataset, we extracted a random subsample of 180 articles from the **XML dataset** (step 9, Fig 1). Two reviewers independently examined each article in the subsample to determine whether or not the data had been deposited in a mentioned repository. The reviewers first examined the context surrounding the mention of the data repository and then, if necessary, the full text of the article. If a determination could not be made by either of these two methods, the reviewers checked the named data repositories for evidence that the data had been deposited. Following the independent reviews, the reviewers met to agree on the final determination for each article.

Analysis of articles with "invisible" datasets

To analyze the 69,657 articles containing "invisible" datasets, we extracted a random sample of 385 articles (confidence interval 95%) for further analysis (step 10, Fig 1). Thirty members of NIH staff were recruited to annotate and analyze the datasets reported in the 385 articles. Annotators were subject experts working in a variety of disciplines including MEDLINE indexers of biomedical literature, biomedical informaticians, physicians, neuroscientists, molecular biologists, librarians, and organizational directors. Each annotator was assigned 25 articles through randomization, and two participants were assigned the same 25 articles—a total of 16 sets—to provide a means to measure the reliability of the counts. One set of annotators only analyzed 10 articles, as this set represented the remaining balance after articles were assigned to other annotators. Each annotator was asked to review his or her assigned articles in their entirety and answer questions related to each dataset described therein. Annotators were instructed to look closely at the methodology section of the paper and any figures or tables to determine how many different measurements were taken. Annotators received a guideline document that included a set of questions to be answered for each assigned article. There was a list of controlled terms for anticipated answers to some of the questions. The guidelines and controlled terms went through several iterations including a pilot study and several internal reviews to improve the clarity of what was being asked and enhance the comparability of the results between annotators. The series of questions are listed in Fig 3.

The categories used to describe the type of data collected in each dataset were developed by the authors of this paper, based on their knowledge of various types of data collected in biomedical research. They do not reflect any particular standard for classifying dataset types. These data types were also informed by an earlier pilot study, and consultation with a variety of stakeholders within the NLM including leadership, indexers, bioinformaticians, and ontologists.

!

1. What category of dataset was used for the research described in the article?

A. New dataset, e.g., lab results or blood pressure measurements after administration of a new drug treatment, new survey results, or mutation analysis of a tumor

B. Existing dataset with modifications or added value, e.g., research using previously collected phenotype data combined with newly collected genotype data

C. Existing dataset as-is, e.g., study using pre-existing survey data to answer a new question

D. None, i.e., no indication that data was created, such as an article about another study that has already been completed.

2. Were live human or animal subjects used in the collection of the data?

A. Yes

B. No

3. If new dataset(s) were created, what were the subject(s) of study (from which or whom the data was collected)?

A. Human (e.g., includes human subjects, tissues or cells)

B. Non-human Animal (e.g. includes animal subjects, tissues or cells)

C. Plant (e.g., includes plant subjects, tissues or cells)

D. Immortalized cell lines (e.g., HeLa, HEK)

E. Bacteria

F. Virus

G. Computational model

H. Other (please provide your best judgment as to the subject of study)

4. If new dataset(s) were created, what type(s) of data were collected?:

A. Image, e.g., two or three dimensions

B. Genetic or genomic, e.g. SNP or genetic insertion/deletion

C. Chemical, e.g., chemical and crystal structures, spectra, reactions and syntheses

D. Biochemical, e.g., structures, functions and interactions of proteins, nucleic acids, carbohydrates, lipids, etc.

E. Electrical (electrophysiological), e.g., EEG or EKG

F. Optical – non-image, e.g., fluorescence signals indicating a biochemical event (non-image)

G. Behavioral – non-questionnaire/survey, e.g., reaction times during a working memory task

H. Computational simulation or model, e.g., computer simulation of fluid mechanical forces in cardiovascular disease development and therapy

I. Magnetic resonance – non-image, e.g., nuclear magnetic resonance (NMR), electron spin resonance (ESR), and electron paramagnetic resonance (EPR)

J. Structural or anatomical, e.g., measures of shape, size and other spatial features of molecules, organelles, cells, tissues, organs or organisms

K. Physiological, e.g., measures of function across interacting parts of the cell, tissue or organism

L. Questionnaire/survey, e.g., collected in epidemiology or health services research, self-report by individuals, etc.

M. Clinical measures, e.g., data assessing quality of care and patient satisfaction

N. Geospatial, e.g., data points related to particular places on the earth

O. Other (please use your best judgment if the type of data is not represented above)

5. If an existing dataset was used, please specify which one:

Examples: a pre-existing survey, MIMIC data, a computational model, previously collected phenotype data, etc.

!

Fig 3. Questions for annotating datasets contained in research articles.

doi:10.1371/journal.pone.0132735.g003

Annotators were asked to populate a spreadsheet with their answers and create a row in the spreadsheet for each dataset found within an article. This procedure provided an opportunity to count how many datasets were created per article, and understand the different types of data that were collected per article. Once annotators completed their 25 articles, the results were returned for review and analysis.

Results

We first summarize the results of our analysis of datasets in known repositories and then present the results of our analysis of the invisible datasets (Fig 1).

Datasets in known repositories

SI Dataset. The use of the SI field identified journal articles for which a dataset had been deposited in a specified repository (step 2, Fig 1). It provided valuable information about the common locations from which data are frequently shared. From the original sample of 113,089 MEDLINE citations, more than 3,500 (3.1%) listed data repositories in the **SI dataset**. The most common repositories where data were deposited were ClinicalTrials.gov, Protein Data Bank, Gene Expression Omnibus, and GenBank (Fig 4).

ACK Dataset. Review of the PMC Acknowledgements field yielded results similar to the SI field search (step 6, Fig 1). The **ACK dataset** (n = 814) articles acknowledged or mentioned a recognized data repository in more than 3,200 instances, for an average of almost 4 datasets

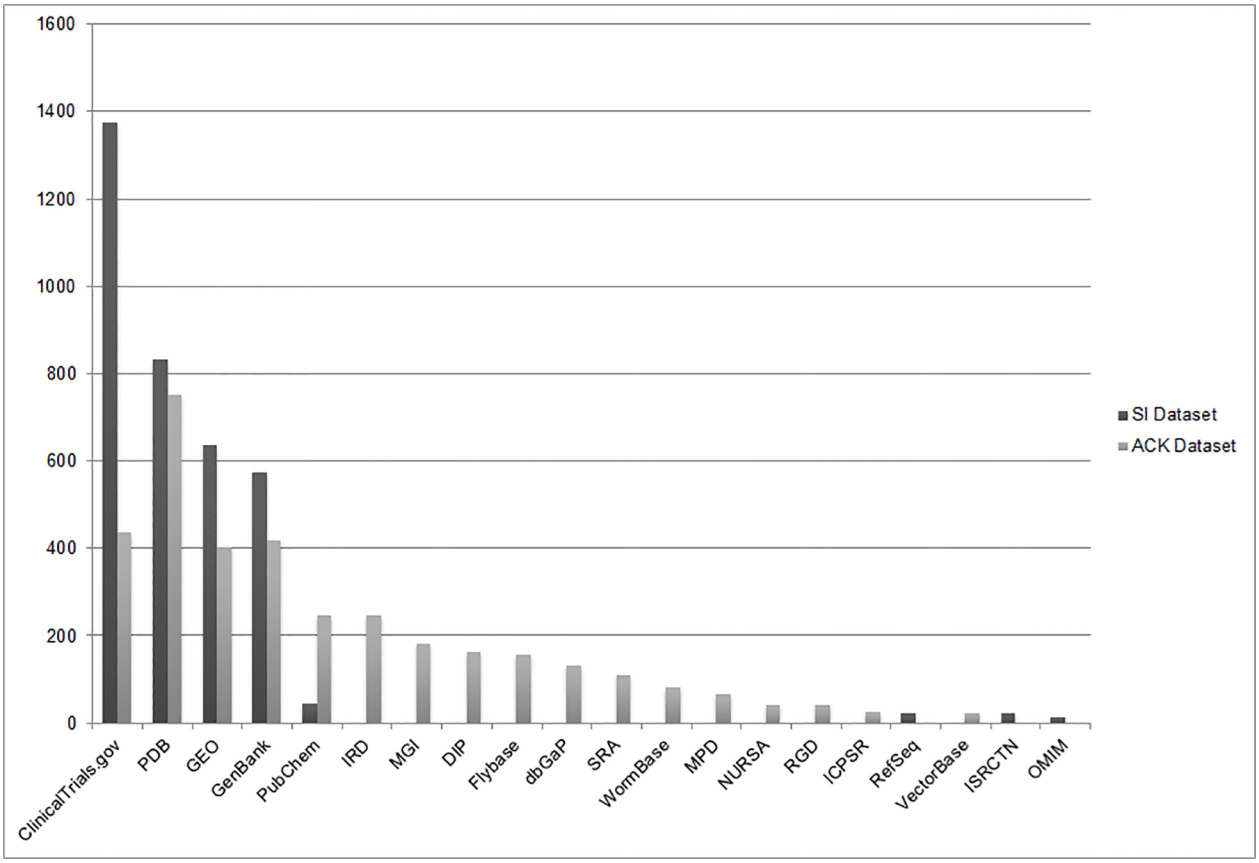


Fig 4. Repositories identified from the PubMed SI field and PMC Acknowledgements where datasets were deposited.

doi:10.1371/journal.pone.0132735.g004

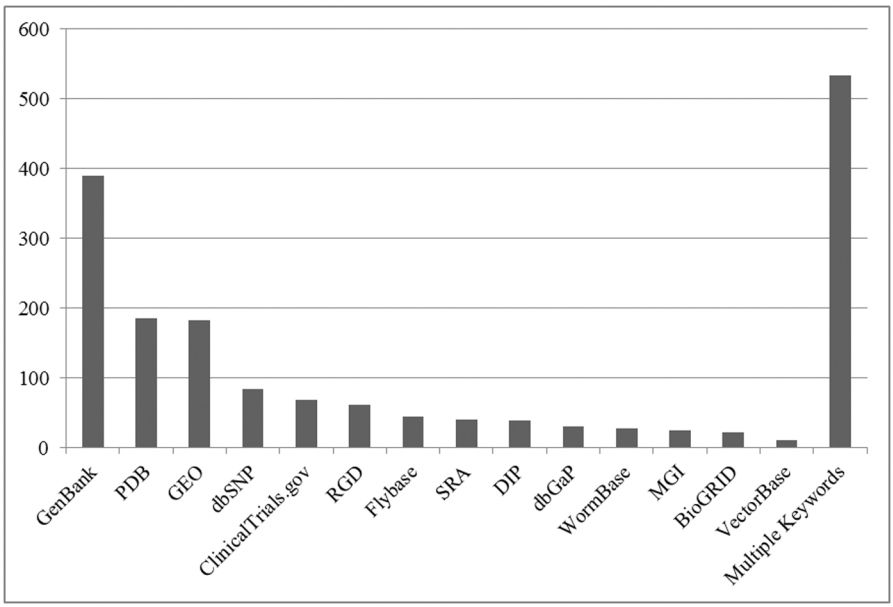


Fig 5. Keywords identified from full-text XML data mining.

doi:10.1371/journal.pone.0132735.g005

per paper. Protein Data Bank, ClinicalTrials.gov, GenBank, and GEO were again the most common repositories where data were being shared (Fig 4). Because the Acknowledgments search identified a wider range of data repositories than were captured in the SI field in 2011, we were able to gain a better understanding of how often other repositories are used. For example, the Influenza Research Database (IRD), Mouse Genome Informatics (MGI) repository, Database of Interacting Proteins (DIP), and Flybase were the most heavily used data repositories beyond the Protein Data Bank (PDB) and databases managed by NLM. This finding provides insight into the frequency of use of these repositories in a given year (Fig 4).

XML Dataset. The final step to identify journal articles that mention a data repository, the XML method, identified 1,825 additional articles that mentioned a data repository somewhere in the text other than the Acknowledgments section (step 8, Fig 1). As noted in the methodology, this analysis was performed on only 10,418 publications from the PMC Open Access Subset, meaning that 17.5% of the articles analyzed were found to mention a dataset. This finding strongly suggests that any future reviews should be expanded beyond the Acknowledgements section to the entire text of an article. The repositories mentioned in the full-text XML aligned with those identified in the **SI** and **ACK datasets**. GenBank, Protein Data Bank, and Gene Expression Omnibus were again the most prominent data repositories mentioned (Fig 5). This method also identified the long tail of deposits in a number of other specialized repositories. One limitation of this approach is that when multiple data repositories were mentioned in the same article, the XML program could not count the separate repositories individually. This caused all articles that mentioned more than one repository to be categorized together. In total, 29% of all the articles mentioned more than one repository.

In estimating how often the mention of a repository in the XML meant that a dataset had actually been deposited in the mentioned repository, the reviewers determined that 33% of the 180 article subsample taken from the **XML dataset** reflected an actual deposit of data into a data repository (step 9, Fig 1). The remainder fell into four main categories: those that used data from a repository (47%); those that mentioned a repository as background information based on previous research (6%); those that discussed a repository as the subject of the article

Table 7. Estimated Number of Articles with a Dataset Stored in a Known Repository.

Procedure taken	Articles Examined	Articles identified	% of Examined Articles
[SI] field (**SI dataset**)	113,089	3,528	3.1%
Molecular Sequence Data MeSH Heading (**MSD dataset**)	109,561	3,460	3.2%
PMC Acknowledgements (**ACK dataset**)	71,910	230	0.3%
Full-text XML (**XML dataset**)	10,418	598	5.7%
Estimated Total	—	—	**12.3%**

doi:10.1371/journal.pone.0132735.t007

(4%); or those that used an ambiguous data repository acronym (10%) (e.g., the acronym RGD for Rat Genome Database is also used to describe arginyl-glycyl-aspartic acid). Extrapolating these findings to our larger set, we estimate that 598 articles in the **XML dataset** (33% of 1,825 articles) would refer to the deposit of data into a named repository. This figure is equal to 5.7% of all the articles examined from the PMC Open Access Subset.

The findings from these various analyses provide a rough estimate of the fraction of NIH-funded research articles that indicate that data were deposited in a known public repository (Table 7). From the larger sample of MEDLINE citations that referenced NIH support in 2011, we found that 3.1% had an SI field (**SI dataset**) that indicated deposit in a known repository (step 2, Fig 1). From the remaining set of articles, another 3.2% had molecular data (**MSD dataset**) that would likely be deposited in GenBank or an equivalent repository (step 3, Fig 1). Of the research articles for which full-text was available, 0.3% included a unique acknowledgement of a data repository (**ACK dataset**) that was discoverable by searching using the [ack] search tag in PMC (step 6, Fig 1). Of articles examined from the PMC Open Access Subset (**XML dataset),** an estimated 5.7% reference the deposit of data into a named repository (step 9, Fig 1). Because the percentages are taken from different subsamples of the larger population of NIH-funded articles in 2011, they cannot be strictly summed. As a rough measure, however, they suggest that an estimated 12.3% (Table 7) of the articles published by NIH-funded investigators in 2011 referred to datasets that are, or may be, stored in a known publicly accessible data repository. The datasets in the remaining 88% of published journal articles were "invisible."

"Invisible" datasets

Our analysis of the 385 journal articles without a discoverable reference to a data repository is summarized in Table 8 and highlights the challenges in defining and counting datasets. Eight of the 30 annotators defined a dataset as consisting of all of the data resulting from an article, irrespective of the different types of data involved (e.g., chemical test results, imaging data). This meant that half of the sets of articles had only one review that was consistent with our proposed methodology, rather than the two desired. As a result, each set of articles that included only one valid review was removed from the final analysis, leaving 8 sets of articles for analysis

Table 8. Summary of Analysis of Invisible Datasets.

Measure	Finding
Average number of datasets per article	All articles reviewed: 3.4 per article
	Articles with two reviews: 2.9 per article
Type of subject	Human subjects: 28.3%
	Non-human animal subjects: 26.1%
New vs. existing data	New datasets: 87%
	Existing datasets: 13%

doi:10.1371/journal.pone.0132735.t008

(step 11, Fig 1). This reduced sample of 200 articles yielded a confidence interval of 84.3% for our subsequent estimates.

For these sets of articles with two reviews, there were substantial differences between annotators in the number of datasets identified and described. Within a set of articles, the average difference between the annotators who identified high and low numbers of datasets was 43%—a significantly high percent difference with respect to the validation of this exercise. While the percentage differences are large, however, the absolute numbers are small, and most pairs differed by only one or two datasets. Only one set of annotators counted widely divergent numbers of datasets in their sample.

The analysis nevertheless provides insight into the number of datasets per article. Considering only the eight sets of articles for which there were two valid reviews, the average number of datasets counted per article was 2.9 (Table 8). When all sets of annotations were evaluated (including those sets where annotators counted only one dataset per article), the average number increased to 3.4 datasets per article, reflecting the fact that some of the publications in the additional sets of articles were reported to contain high numbers of datasets. An average of between 2.9 and 3.4 datasets per article aligns with the estimated four datasets per paper we found in our methods for identifying datasets within the **ACK dataset**.

There was greater consistency between pairs of annotators in identifying data from human subjects and live non-human animals. The average percentage of articles identified as reporting research involving human subjects was 28.3%, and the percentage identified as involving non-human animals was 26.1% (Table 8).

The last phase of analysis determined how much of the data that was used in the course of NIH-funded research published in 2011 was new data, and how much was pre-existing data. We counted the percentage of new versus pre-existing data for each set of articles and then calculated the total percentage from all. Annotators were very consistent in making this determination. Combining results from all annotators, we estimate that 87% of the articles involved the collection of new data, and 13% involved the analysis of pre-existing data (Table 8). While new data were collected for purposes of the research reported in the article, pre-existing data included data from previously conducted clinical trials (e.g., reanalysis of the clinical trial data) or surveys (e.g., at local, regional, or national level), among other sources. Some articles made use of both new and pre-existing data.

Annotators were inconsistent in their ability to assign data types from our controlled list of categories to datasets found within articles. Few annotators chose to use the controlled list; most preferred to use the "Other" option to describe the datasets they found, highlighting the difficulty in establishing a suitable classification for biomedical data types (Fig 3).

Discussion

This study is an initial step toward estimating the additional resources and infrastructure that will be needed to support expanding mandates to make data resulting from NIH-funded and other research available for use by researchers and the public. Methods for discoverability, access, and citation will need to be able to scale in a cost-effective manner if they are to include all such datasets used in published NIH-funded research studies, let alone all datasets used in published biomedical research regardless of funder. It is likely that an increasing number of biomedical datasets will be deposited in general purpose repositories (e.g., Dryad, Figshare), which will affect strategies for enhancing discoverability and access. Data citations are still uncommon and not frequently used by the scientific community. One study found that 88.1% of datasets within the Data Citation Index remained uncited [25]. Another study found that even a national data center rarely could identify formal citations of their data [26].

Without the ability to link and connect research datasets across multiple platforms, discovery and access will remain an issue.

Our results suggest that datasets referenced in only about 12% of articles reporting NIH-funded research in 2011 were (or were eligible to be) deposited in a known, publicly accessible data repository. Our estimates echo findings from previous studies that indicate a large portion of datasets are not shared. In one study, only 9% of articles from high-impact journals deposited their full dataset (including raw data) online [12]. Another study found that in their sample not a single study cited a dataset with a unique identifier, therefore providing no indication that the data are shared anywhere [13]. Our analysis also found that an expected 858 articles (47% of 1,825 articles) in the **XML dataset** mention the use of data from a known repository (rather than a deposit into a repository). This figure is equal to 8.2% of the articles we examined from the PMC Open Access Subset and provides a measure of the level of reuse of data from known repositories.

Our findings also help characterize the invisible datasets from biomedical research that are not deposited in a known repository. We found 69,657 articles that were published in 2011 and reported on NIH-funded research but did not indicate that data had been deposited in a known repository. We estimate that the research described in these articles used an average of 2.9 to 3.4 datasets per article. These figures mean that approximately 200,000 to 235,000 datasets were used in NIH-funded research published in 2011 but not deposited in one of the well-known public repositories for specific categories of biomedical data (e.g., GEO, GenBank, Protein DataBank, ClinicalTrials.gov).

Given that as many as 88% of biomedical research datasets may not currently be deposited in a well-known, public data repository, the problem of improving the discoverability of biomedical datasets remains significant. The basic challenge is therefore deciding which data are most worthy of the additional resources and effort that will inevitably be required to make them readily discoverable and accessible. Arguably, useful and manageable data discovery systems should focus on datasets that show potential for reuse or that point to significant findings so that underlying data should be available to others to assess validity and accuracy. A strong case can be made that datasets derived from live subjects have a particularly high priority for discoverability and accessibility. Making such datasets readily available could reduce the need to expose additional live subjects to potential risks and, in the case of human subjects, help to meet the ethical obligation to ensure that their participation in research studies adds to scientific knowledge.

Another point to consider is *how* data can be best discovered, accessed, and understood. Our review of datasets suggests that while some datasets (such as those stored in known repositories like ClinicalTrials.gov, GenBank, and Protein DataBank) can stand on their own and serve as resources for other investigators, many datasets may have limited utility outside the study for which they were collected. These datasets may be meaningful only when considered alongside other datasets collected for the same study and in conjunction with the journal article that summarizes them. This observation has significant implications for data discovery and storage, because it suggests that in some cases the preferred discovery tool may be the publication where the datasets are described rather than a separate mechanism that would find and retrieve them individually and independently. This argument has already been addressed within the scientific community, with some calling for an advanced publication where the underlying data can be extracted directly from the paper [27–30]. Nanopublications are another development that shed light on providing context for datasets pulled from a scientific paper; these abridged data publications provide narrative descriptions of data pulled directly from an article [31]. It is our belief that including dataset metadata summaries within the published article may be an efficient way to promote the discovery of these datasets.

Further work is also needed to determine how to define a dataset. As evidenced by the lack of consistency between annotators with respect to the number of datasets they identified, there are differences in perceptions of what constitutes a dataset and in how well data collected or used in a research study are described in journal articles. Depending on one's perspective, a single dataset could be: all of the data that is collected or used in a study; all data collected at a specific time within a study; pre- or post-intervention; a discrete type of data from a specific diagnostic device; or even every individual measurement reported in a research article. Data access and sharing requirements must clearly define a dataset to outline expectations of what researchers will be required to share and submit and what will be available to potential users. Requirements are likely to vary depending on the use (or reuse) cases for different types of data.

Data creation and analysis pipelines raise additional questions about how data should be described and at what point along the pipeline. Collected data go through multiple processing transformations during analysis. As a simple example, an image may be collected of a cell (e.g., an optical image); that image may be analyzed by measuring the size of certain structures in the cell (e.g., numerical data); the numerical/structural data from multiple cells may be aggregated for analysis (e.g., to compare the size of structures in treated versus untreated cells). Results of that analysis may be shown in a table or a graph, perhaps showing trends in size of the structures in the treated and untreated cells over time. For a researcher interested in reproducing the research, the basic imaging data may be of most interest, but such a researcher might also want or need to know how the data were reduced and need access to associated data processing algorithms. For a researcher interested in comparing results across studies, the more processed data may be of most interest. For a researcher interested in reusing the data, the data from a particular point along the data processing pipeline might be most useful. Providing data at each step along the pipeline might prove to be onerous or overly complex for data generators and those who want to make use of the data.

Any system for data discovery and access must describe data in a way that will be useful for those researchers, health professionals or members of the public who are interested in reviewing biomedical data. An examination of a variety of metadata schemas [32–34] and the metadata employed in existing NIH data repositories indicates that the baseline description of datasets in current repositories does not differ greatly from descriptive metadata for journal articles or archival objects. However, we are not aware of strong evidence that the current metadata schemes applied to biomedical datasets either do or do not meet the needs of researchers seeking data to reuse. There has been considerable discussion about including enriched metadata to make data more discoverable in the context of a data publication to provide detailed metadata and description of individual datasets [35, 36]. Some research has begun to examine the quality of metadata used in scientific data repositories [37], but more research is needed to determine what metadata would enable efficient discovery of various types of data. Analysis of current use patterns of existing repositories that accommodate disparate datasets may shed light on what types of data and descriptive metadata are most useful.

Determining which types of biomedical data have the highest reuse value, how to describe them usefully and cost-effectively, and how to make them accessible in a sustainable way are key challenges for the NIH and its recently established Data Discovery Index Coordination Consortium [38] as they move forward to make biomedical big data more discoverable, accessible, and citable.

Conclusion

These findings represent a first look into the landscape of NIH-funded data. An understanding of the varying types of data that are created throughout the course of biomedical research and

the knowledge that a substantial amount of new data is created per article in a given year will help to inform efforts to improve the discoverability and accessibility of digital biomedical research data. Differences in perspective encountered among participants in the study suggest that the creation of data discovery tools for biomedical research data will not be straightforward. Decisions will have to be made as to what data will be selected for description and careful consideration will need to be given to identifying how to describe datasets derived from NIH-funded research.

Acknowledgments

Investigators (institution and location) in the NIH Big Data Annotator Group include (in alphabetical order): Swapna Abhyankar (National Library of Medicine, Bethesda, MD), Olubumi Akiwumi (Oregon Health & Science University, Portland, OR), Olivier Bodenreider (National Library of Medicine, Bethesda, MD), Sally Davidson (National Library of Medicine, Bethesda, MD), Dina Demner Fushman (Library of Medicine, Bethesda, MD), Tracy Edinger (Kaiser Permanente, Portland, OR), Greg Farber (National Institute of Mental Health, Bethesda, MD), Karen Gutzman (Bernard Becker Medical Library, Chicago, IL), Mary Ann Hantakas (National Library of Medicine, Bethesda, MD), Preeti Kochar (National Library of Medicine, Bethesda, MD), Jennie Larkin (National Heart Lung and Blood Institute, Bethesda, MD), Peter Lyster (National Institute of General Medical Sciences, Bethesda, MD), Matt McAuliffe (Federal Interagency Traumatic Brain Injury Research Informatics System, Bethesda, MD), Shari Mohary (National Library of Medicine, Bethesda, MD), Helen Ochej (National Library of Medicine, Bethesda, MD), Olga Printseva (National Library of Medicine, Bethesda, MD), Oleg Rodionov (National Library of Medicine, Bethesda, MD), Laritza Rodriguez (National Library of Medicine, Bethesda, MD), Suzy Roy (National Library of Medicine, Bethesda, MD), Susan Schmidt (National Library of Medicine, Bethesda, MD), Sonya Shooshan (National Library of Medicine, Bethesda, MD), Matthew Simpson (National Library of Medicine, Bethesda, MD), Corinn Sinnot (National Library of Medicine, Bethesda, MD), Samantha Tate (National Library of Medicine, Bethesda, MD), Janice Ward (National Library of Medicine, Bethesda, MD), Melissa Yorks (National Library of Medicine, Bethesda, MD).

We gratefully acknowledge Lori Klein, National Library of Medicine for her assistance with the preparation of the References list.

This research was supported by the Intramural Research Program of the U.S. National Institutes of Health, National Library of Medicine (NLM) and in part by an appointment to the NLM Associate Fellowship Program sponsored by the National Library of Medicine and administered by the Oak Ridge Institute for Science and Education.

Author Contributions

Conceived and designed the experiments: KBR JRS MFH LSK JGM BLH. Performed the experiments: KBR JRS MFH LSK JGM BLH NBDAG. Analyzed the data: KBR JRS MFH LSK JGM BLH. Wrote the paper: KBR JRS MFH LSK JGM BLH.

References

1. Chan AW, Song F, Vickers A, Jefferson T, Dickersin K, Gøtzsche PC, et al. Increasing value and reducing waste: addressing inaccessible research. Lancet. 2014 Jan 18; 383(9913):257–66. doi: 10.1016/S0140-6736(13)62296-5 Epub 2014 Jan 8. PMID: 24411650. Accessed 20 Feb 2015.
2. Névéol A, Wilbur WJ, Lu Z. Extraction of data deposition statements from the literature: a method for automatically tracking research results. Bioinformatics. 2011 Dec 1; 27(23):3306–12. doi: 10.1093/bioinformatics/btr573 Epub 2011 Oct 13. PMID: 21998156; PubMed Central PMCID: PMC3223368. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3223368/. Accessed 9 June 2015.

3. OECD. [Paris]: Organization of Economic Co-operation and Development. Open science; [accessed 2015 Jun 10]. Available: http://www.oecd.org/sti/outlook/e-outlook/stipolicyprofiles/interactionsforinnovation/openscience.htm. Accessed 10 Jun 2015.
4. EU Framework Programme for Research and Innovation. Guidelines on open access to scientific publications and research data in Horizon 2020. Version 16. [place unknown]: European Commission; 2013 Dec. 14 p. Available: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf. Accessed 4 Mar 2015.
5. European Research Council, Scientific Council. Open access guidelines for research results funded by the ERC. [place unknown]: European Research Council; revised 2014 Dec. 3 p. Available: http://erc.europa.eu/sites/default/files/document/file/ERC_Open_Access_Guidelines-revised_2014.pdf. Accessed 15 Mar 2015.
6. Tri-Agency open access policy on publications. [Ottawa (ON)]: Government of Canada, Public Works and Government Services Canada Publishing and Depository Services; 2015 [modified 2015 Feb 27; accessed 2015 Mar 12]. [about 3 p.]. Available: http://www.science.gc.ca/default.asp?lang=En&n=F6765465-1
7. Holdren JP (Director, Office of Science and Technology Policy, Executive Office of the President, Washington, DC). Increasing access to the results of federally funded scientific research. Memorandum to: Heads of Executive Departments and Agencies. 2013 Feb 22. 6 p. Available: http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf. Accessed 1 Mar 2014.
8. National Institutes of Health plan for increasing access to scientific publications and digital scientific data from NIH funded scientific research. [Bethesda (MD)]: U.S. Department of Health and Human Services, National Institutes of Health; 2015 Feb. 44 p. Available: http://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf. Accessed 12 Feb 2015.
9. National Institutes of Health (US). Bethesda (MD): U.S. Department of Health and Human Services, National Institutes of Health (US); NIH budget; [reviewed 2015 Jan 29; accessed 2015 Mar 19]; [about 3 screens]. Available: http://www.nih.gov/about/budget.htm. Accessed 19 Mar 2015.
10. Big data to knowledge (BD2K). Bethesda (MD): U.S. Department of Health and Human Services, National Institutes of Health (US); 2012 [last updated 2015 Jun 1; accessed 2015 Jun 9]. Available: https://datascience.nih.gov/bd2k
11. Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, et al. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc. 2014 Nov-Dec; 21(6):957–8. doi: 10.1136/amiajnl-2014-002974 Epub 2014 Jul 9. PMID: 25008006; PubMed Central PMCID: PMC4215061. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4215061/. Accessed 15 Mar 2015.
12. Alsheikh-Ali AA, Qureshi W, Al-Mallah MH, Ioannidis JP. Public availability of published research data in high-impact journals. PLoS One. 2011; 6(9):e24357. doi: 10.1371/journal.pone.0024357 Epub 2011 Sep 7. PMID: 21915316; PubMed Central PMCID: PMC3168487. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3168487/. Accessed 20 Feb 2015.
13. Mooney H, Newton MP. The anatomy of a data citation: discovery, reuse, and credit. J Librariansh Sch Commun. 2012: 1(1):eP1035. Available: doi: 10.7710/2162-3309.1035. Accessed 20 Feb 2015.
14. Belter CW. Measuring the value of research data: a citation analysis of oceanographic data sets. PLoS One. 2014 Mar 26; 9(3):e92590. doi: 10.1371/journal.pone.0092590 eCollection 2014. PMID: 24671177; PubMed Central PMCID: PMC3966791. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3966791/. Accessed 20 Feb 2015.
15. Piwowar HA, Carlson D, Vision TJ. Beginning to track 1000 datasets from public repositories into the published literature. Proc Am Soc Info Sci Technol. 2011 [published online 2012 Jan 11; accessed 2013 May 20]; 48(1):1–4. doi: 10.1002/meet.2011.14504801337 Available: http://onlinelibrary.wiley.com/doi/10.1002/meet.2011.14504801337/abstract Poster.
16. Ariño A. Approaches to estimating the universe of natural history collections data. Biodivers Inf. 2010; 7 (2):81–92. Available: doi: 10.17161/bi.v7i2.3991. Accessed 20 Feb 2015.
17. Ross JS, Tse T, Zarin DA, Xu H, Zhou L, Krumholz HM. Publication of NIH funded trials registered in ClinicalTrials.gov: cross sectional analysis. BMJ. 2012 Jan 3; 344:d7292. doi: 10.1136/bmj.d7292 PMID: 22214755; PubMed Central PMCID: PMC3623605. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3623605/. Accessed 6 Nov 2014.
18. Vines TH, Albert AY, Andrew RL, Débarre F, Bock DG, Franklin MT, et al. The availability of research data declines rapidly with article age. Curr Biol. 2014 Jan 6; 24(1):94–7. doi: 10.1016/j.cub.2013.11.014 Epub 2013 Dec 19. PMID: 24361065. Accessed 20 Feb 2015]

19. PubMed help. Bethesda (MD): U.S. National Library of Medicine, National Center for Biotechnology Information; 2005 -. Secondary Source ID; [2 paragraphs]. Available: http://www.ncbi.nlm.nih.gov/books/NBK3827/#pubmedhelp.Secondary_Source_ID_SI. Accessed 12 Jun 2014.
20. PMC help. Bethesda (MD): U.S. National Library of Medicine, National Center for Biotechnology Information; 2005 -. Acknowledgements [ACK]; [1 paragraph]. Available: http://www.ncbi.nlm.nih.gov/books/NBK3825/#pmchelp.Acknowledgements_ACK. Accessed 29 Jul 2014.
21. Hinchliff CE, Smith SA. Some limitations of public sequence data for phylogenetic inference (in plants). PLoS One. 2014 Jul 7; 9(7):e98986. doi: 10.1371/journal.pone.0098986 eCollection 2014. PMID: 24999823; PubMed Central PMCID: PMC4085032. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4085032/. Accessed 3 Sep 2014.
22. Trans-NIH Biomedical Informatics Coordinating Committee (BMIC). Bethesda (MD): National Institutes of Health, U.S. National Library of Medicine; 2013 Jan 4. NIH data sharing repositories; 2013 Jan 23. Available: http://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html. Accessed 2 Aug 2013.
23. National Library of Medicine. Bethesda (MD): National Institutes of Health (US), National Library of Medicine; 1993. MEDLINE PubMed XML element descriptions and their attributes; 2005 Dec [last modified 2012 Dec; accessed 2013 Aug 4]. Available: http://www.nlm.nih.gov/bsd/licensee/elements_descriptions.html
24. PMC. Bethesda (MD): U.S. National Library of Medicine, National Center for Biotechnology Information; 2000. PMC open access subset; [2013; updated 2014 Jan 13; accessed 2014 Dec 10]. Available: http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/.
25. Robinson-Garcia N, Jimenez-Contreras E, Torres-Salinas D. Analyzing data citation practices using the Data Citation Index. J Assoc Inf Sci Technol. 2015 Jun 1:[12 p.]. Available: http://onlinelibrary.wiley.com/doi/10.1002/asi.23529/abstract doi: 10.1002/asi.23529. Also available from: http://arxiv.org/abs/1501.06285. Accessed 9 Jun 2015.
26. Parson MA, Duerr R, Minster JB. Data citation and peer review. EOS. 2010 Aug 24; 91(34):297–8. doi: 10.1029/2010EO340001 Available: http://onlinelibrary.wiley.com/doi/10.1029/2010EO340001/full. Accessed 20 Feb 2015.
27. Callaghan S. Preserving the integrity of the scientific record: data citation and linking. Learn Publ. 2014; 27:S15–S24. doi: 10.1087/20140504 Available: http://www.ingentaconnect.com/content/alpsp/lp/2014/00000027/00000005/art00004. Accessed 18 Feb 2015.
28. Lynch C. The shape of the scientific article in the developing cyberinfrastructure. CTWatch Q. 2007 Aug; 3(3):5–10. Available: http://www.ctwatch.org/quarterly/articles/2007/08/the-shape-of-the-scientific-article-in-the-developing-cyberinfrastructure/. Accessed 12 Feb 2013.
29. Lindberg DA. Research opportunities and challenges in 2005. Methods Inf Med. 2005; 44(4):483–6. PMID: 16342914. Available: http://methods.schattauer.de/en/contents/archivestandard/issue/685/manuscript/504/show.html. Accessed 28 Apr 2015.
30. Thoma GR, Ford G, Antani S, Demner-Fushman D, Chung M, Simpson M. Interactive publication: the document as a research tool. Web Semant. 2010 Jul 1; 8(2–3):145–150. PMID: 20657757; PubMed Central PMCID: PMC2908409. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2908409/. Accessed 2 Mar 2015.
31. Mons B, van Haagen H, Chichester C, Hoen PB, den Dunnen JT, van Ommen G, et al. The value of data. Nat Genet. 2011 Mar 29 [accessed 2015 Feb 20]; 43(4):281–3. doi: 10.1038/ng0411-281 PMID: 21445068.
32. DataCite. London: DataCite; [accessed 2014 Aug 11]. DataCite Metadata Schema Repository; [last updated 2013 Jul 24; accessed 2014 Aug 11]. Available: http://schema.datacite.org/.
33. Dryad Digital Repository. Durham (NC): Dryad. 2008 Jan—. Metadata profile: Dryad metadata application profile (schema); [last modified 2013 Feb 27; accessed 2014 Aug 3]. Available: http://wiki.datadryad.org/Metadata_Profile
34. W3C. [place unknown]: World Wide Web Consortium; c2014. Data Catalogue Vocabulary (DCAT); 2014 Jan 16. Available: http://www.w3.org/TR/vocab-dcat/. W3C recommendation. Accessed 7 Feb 2014.
35. Chavan V, Penev L. The data paper: a mechanism to incentivize data publishing in biodiversity science. BMC Bioinformatics. 2011; 12 Suppl 15:S2. PMID: 22373175; PubMed Central PMCID: PMC3287445. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3287445/. Accessed 20 Feb 2015.
36. Costello MJ, Michener WK, Gahegan M, Zhang ZQ, Bourne PE. Biodiversity data should be published, cited, and peer reviewed. Trends Ecol Evol. 2013 Aug [accessed 2015 Feb 20]; 28(8):454–61. doi: 10.1016/j.tree.2013.05.002 Epub 2013 Jun 5. PMID: 23756105.
37. Rousidis D, Garoufallou E, Balatsoukas P, Sicilia MA. Metadata for Big Data : a preliminary investigation of metadata quality issues in research data repositories. Inf Serv Use. 2014; 34(3–4):279–86. doi: 10.3233/ISU-140746. Accessed 12 Dec 2014.

38. Big data to knowledge (BD2K). Bethesda (MD): U.S. Department of Health and Human Services, National Institutes of Health (US); 2012 [last updated 2015 Jun 1]. Data Discovery Index Coordination Consortium (DDICC) (University of California, San Diego). BioCADDIE: Biomedical and healthcare data discovery and indexing engine center; [about 1 p.]. Available: https://datascience.nih.gov/sites/default/files/bd2k/docs/DDIC.pdf. Accessed 9 Jun 2015.